Research Article

Forward Variable Selection Improves the Power of Random Forest for High-Dimensional Micro Biome Data

Tung Dang[1*], Hirohisa Kishino[1,2]

[1]Department of Agricultural and Environmental Biology, The University of Tokyo, Tokyo, Japan

[2]The Research Institute of Evolutionary Biology, Tokyo, Japan

[*]**Corresponding Author:** Tung Dang, Department of Agricultural and Environmental Biology, The University of Tokyo, Tokyo, Japan.

Received: 04 February 2022; **Accepted:** 17 February 2022; **Published**: 23 February 2022

Citation: Tung Dang, Hirohisa Kishino. Forward Variable Selection Improves the Power of Random Forest for High-Dimensional Micro Biome Data. Journal of Cancer Science and Clinical Therapeutics 6 (2022): 87-105.

Abstract

Random forest (RF) captures complex feature patterns that differentiate groups of samples and is rapidly being adopted in microbiome studies. However, a major challenge is the high dimensionality of microbiome datasets. They include thousands of species or molecular functions of particular biological interest. This high dimensionality significantly reduces the power of random forest approaches for identifying true differences and functional characterization. The widely used Boruta algorithm iteratively removes features that are proved by a statistical test to be less relevant than random probes. We developed a massively parallel forward variable selection algorithm and coupled it with the RF classifier to maximize the predictive performance. The forward variable selection algorithm adds new variable to a set of selected variables as far as the prespecified criterion of predictive power is improved. At each step, the parameters of random forest are optimized. We demonstrated the performance of the proposed approach, which we named RF-FVS, by analyzing two published datasets from large-scale case-control studies: (i) 16S rRNA gene amplicon data for Clostridioides Difficile Infection (CDI) and (ii) shotgun metagenomics data for human colorectal cancer (CRC). The RF-FVS approach further screened the variables that the Boruta algorithm left and improved the accuracy of the random forest classifier from 81% to 99.01% for CDI and from 75.14% to 90.17% for CRC. Valid variable selection is essential for the analysis of high-dimensional microbiota data. By adopting the Boruta algorithm for pre-screening of the variables, our proposed RF-FVS approach improves the accuracy of random forest significantly with minimum increase of computational burden. The procedure can be used to identify the functional profiles that differentiate samples between different conditions.

1. Introduction

A microbiome is the full collection of genes of all microbes in a community; for example, all bacteria in a sample from the gut of a healthy individual or from an individual with a disease. Identifying difference of microbiome compositions between two or more groups is one of the most important purposes of microbiome studies [1, 2]. High-throughput sequencing technologies have allowed the microbiome composition and function in different environments to be quantified correctly [3, 4]. Several marker identification methods have been developed for applications in microbiome studies. The standard statistical approaches, such as Kruskal-Wallis (KW) test with the Benjamini–Hochberg False Discovery Rate (FDR) correction [5] or blocked (univariate) Wilcoxon tests [6], measure taxon relative abundances, analyze within- and between-sample diversity (α and β diversity, respectively), and perform classical hypothesis testing. Machine learning technology has been applied in microbiome studies, especially for predicting specific diseases and supporting medical diagnosis [7, 8]. Because Random Forest (RF) captures the complex feature patterns that differentiate groups of samples [9, 10], it is rapidly being adopted for the analysis of microbiome data. The RF algorithm is a modification of bagging that aggregates a large collection of decision trees [11]. A main step in building an ensemble of decision trees is to perform random sampling of the available features to generate different subspaces of features at each node of each unpruned decision tree. This strategy can produce better estimation performances than a single decision tree because each tree estimator has low bias but high variance, whereas a bias-variance trade-off is achieved by the bagging process of RFs. RF methods have been applied successfully to genetic and microbiome data [12-14]. It is anticipated that RF methods and implemented importance measures will help in the identification of microbiome species that can be used to distinguish diseased and non-diseased samples. Identifying a core set of the most significant microbial species is of high interest, not only for diagnosis of certain diseases but also to gain valuable insights into the biological functionality and mechanisms of these species.

However, the performance and diversity of decision trees in the ensemble significantly influence the performance of RF algorithms. The generalization error for RFs involves measures of how accurate the individual classifiers are and their interdependence. Therefore, the high dimensionality problems of microbiome datasets pose a number of challenges. For example, microbiome datasets tend to contain a large number of microbiome species whose functions may not be related to the disease of interest. Common random sampling methods may select a sizeable number of subspaces that do not include the informative microbiome species and functions. As a consequence, the decision trees generated from these subspaces will have reduced average strength, thereby increasing the error bounds for the RF algorithm. A number of different approaches have been proposed to identify important variables that could improve the performance of RF algorithms. For example, the Boruta algorithm [15] was proposed to identify a set of relevant features using an RF classification algorithm that iteratively removes the variables using a statistical test. These relevant features are different from the objective of relevant and also non-redundant feature subsets. Moreover, a standard permutation test [16] was proposed to estimate the distribution of measured importance values of the RF algorithm for each predictor variable by repeatedly permuting the variable and randomly shuffling the data values so that the original association between the response and predictor variables was destroyed. A high value of the permutation importance of the predictor variable indicates high significant association to the response. However, the sizes of the selected subsets of features are still large in the high-dimensional microbiome database, the power of RF algorithm is not significantly improved, and it is difficult to interpret the selected features.

Although a number of different algorithms and tools have been developed for microbiome analysis [7], effective approaches require a lot of possible combinations of variables, which exponentially increases the computational burden as the number of involved features increases. Even though a small number of machine learning methods, including the RF algorithm, can be easily parallelized, building prediction models for thousands of microbiome species and functions can be very time consuming.

In this study, we propose a novel procedure that tackles the challenges described above. The core of our procedure is an RF classifier coupled with forward variable selection (RF-FVS), which selects a minimal-size core set of microbial species or functional signatures to maximize the predictive performance of the RF classifier. To reduce the computational cost, we designed a parallelized algorithm and integrated a prescreen algorithm. To examine the performance of the RF-FVS approach, we analyzed two empirical datasets from large-scale case control studies. One is a published case-control 16S rRNA gene amplicon sequencing gut microbiome dataset with 3347 Operational Taxonomic Units (OTUs) for Clostridioides groups: 89 individuals with CDI (cases), 89 with diarrhea who tested negative for CDI (diarrheal controls), and 155 non-diarrheal controls. The other is a fecal shotgun metagenomic dataset [18] that included 290 samples from tumor-free. Controls and 285 samples from individuals with Colorectal Cancer (CRC). The RF-FVS serves as the basis for the integrated analysis of microbiome data, detecting significant species in a phylogenetic tree [19] and predicting functional capabilities of microbial communities based on 16S rRNA datasets [20]. To estimate the dependence of the pipeline on the quality of the microbiome functional profile data, we analyzed the predicted functional profiles obtained from the 16S rRNA dataset, which were approximately 85% accurate, and the functional profiles obtained from the shotgun metagenomic dataset. We found that our RF-FVS performed well even for the predicted functional profile data.

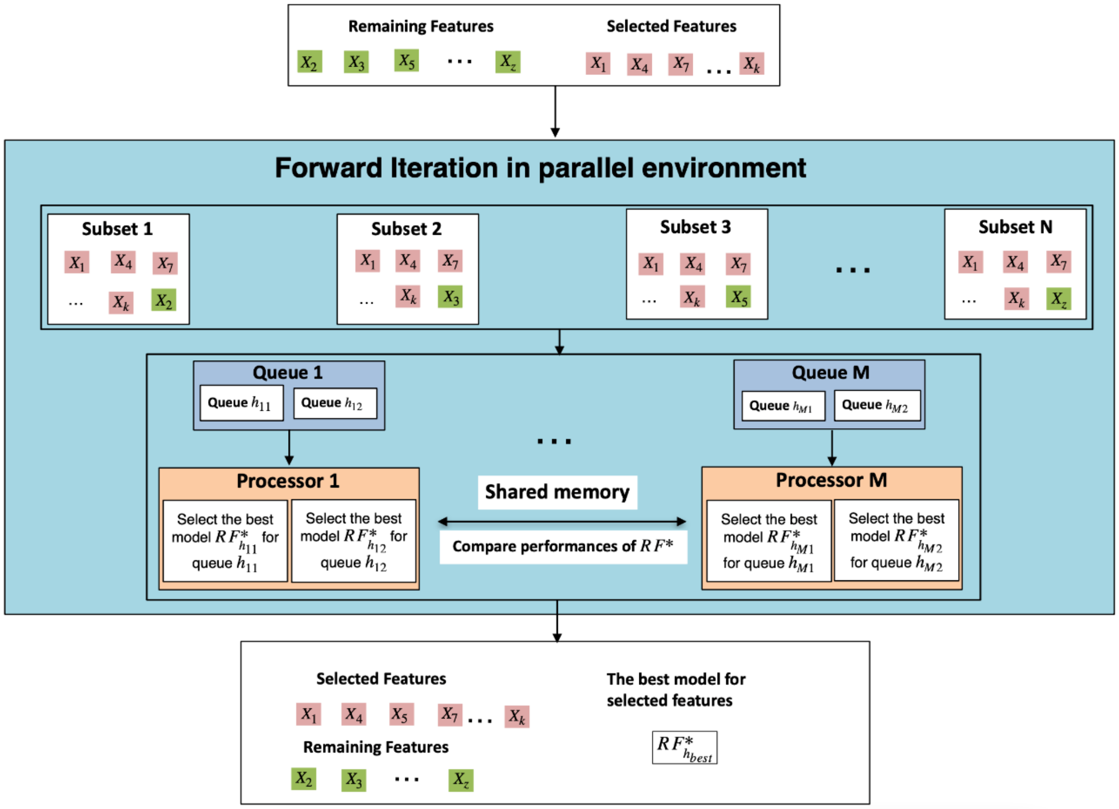
Remaining Features
X_2 X_3 X_5 ... X_z
Selected Features
X_1 X_4 X_7 ... X_k
Forward Iteration in parallel environment
Subset 1
X_1 X_4 X_7 ... X_k X_2
Subset 2
X_1 X_4 X_7 ... X_k X_3
Subset 3
X_1 X_4 X_7 ... X_k X_5
...
Subset N
X_1 X_4 X_7 ... X_k X_z
Queue 1
Queue h_{11}
Queue h_{12}
...
Queue M
Queue h_{M1}
Queue h_{M2}
Processor 1
Select the best model $RF^*_{h_{11}}$ for queue h_{11}
Select the best model $RF^*_{h_{12}}$ for queue h_{12}
Shared memory
Compare performances of RF^*
Processor M
Select the best model $RF^*_{h_{M1}}$ for queue h_{M1}
Select the best model $RF^*_{h_{M2}}$ for queue h_{M2}
Selected Features
X_1 X_4 X_5 X_7 ... X_k
Remaining Features
X_2 X_3 ... X_z
The best model for selected features
$RF^*_{h_{best}}$

Figure 1: Massively parallel forward variable selection algorithm for the Random Forest (RF) classifier. The number of queues depends on the number of CPU cores available in the computer system.

2. Materials and Methods

2.1 Forward variable selection for random forests

The RF approach is an ensemble method that combines a large number of individual binary decision trees. Two main randomization procedures have been implemented to reduce variance of individual decision trees, deliver diversity amongst decision trees, and thus improve prediction accuracy. First, randomly selected training samples for each of the individual trees are applied to construct sufficiently diverse trees. Second, at each node within a tree, a set of randomly chosen candidate predictor variables is identified for the split. However, random feature subspace sampling may not be a good strategy to deal with high-dimensional data because a large proportion of the features may not be informative of the class of an object in the high-dimensional data. If a random sampling strategy is implemented to select the subset of eligible features at each node, almost all the subsets are likely to contain a large number of non-informative features. For example, the 16S rRNA gene amplicon data for CDI [17] that we used to evaluate the performance of our proposed approach contained a total of 3347 microbiome species, but only 96 of the species were informative. Therefore, if a subset of species, which is usually the square root of the total number of species, is selected by resampling randomly at any node within the decision tree, the mean number of informative species selected at each node will be two. Therefore, individual decision trees built using such nodes will have low accuracy and the performance of the RF algorithm will suffer. In our approach, we used forward variable selection to identify a small number of informative variables to improve the performance of individual decision trees in an ensemble.

A key idea behind our algorithm was to divide the total number of variables into two groups, a remaining variables group and a selected variables group. We started with an empty group for the selected variables. At each step, a variable from the remaining variables group was added to the selected variables group such that the specified criterion was improved (i.e., area under the Receiver Operating Characteristic [ROC] curve [AUC], a weighted average of the precision and recall [F1 score] or predictive accuracy). Model selection for microbial signature identification also can be performed using our RF- FVS algorithm. At each forward iteration, given the selected variables, the randomized parameter optimization algorithm for RF implements a randomized search over parameters, where each setting is sampled from a distribution over possible parameter values. Thus, the best RF model is specified by these selected variables. Moreover, a high-speed computational strategy based on multi-processing architecture was developed to parallelize the forward variable selection algorithm at the single machine level and thus significantly reduce runtimes. Another key idea behind our algorithm was to create many subsets of variables, so that each subset had one of the variables from the remaining variables group added to the selected variables group. Because of the high dimensionality problems of microbiome data, the number of these subsets is usually significantly larger than the number of processors in a single computer system. Our solution was to create queues so that subsets are assigned randomly and each processor runs the computational processes from its own privately prepared queue. (Figure 1) shows how all computational burdens for searching important feature relevance are appropriately decomposed, so that they can be computed in a parallel environment. The RF model with the highest accuracy value for each subset of features is selected by a specific processor. Processors in symmetric multiprocessing communicate with each other through shared memory architecture to decide only the best feature among multiple candidates that should be added to the selected features. The algorithm stops when there are no additional variables that improve the current optimization parameters or when the maximum number of components to be included in the group of selected variables is achieved. The main parameters of the RF classifier that were optimized in our algorithm, were number of trees in the forest, maximum depth of the tree, minimum number of samples required to split an internal node, minimum number of samples required to be at a leaf node, and number of features to consider when looking for the best split.

2.2 Functional gene enrichment analysis

Functional profiles were predicted from the 16S rRNA gene data for CDI using Tax4Fun. which was developed to analyze the enrichment of functional genes of microbiomes [20, 21]. The output from the QIIME software application with a SILVA database extension (SILVA 119) [22] was used to pre-process raw data for Tax4Fun. Tax4Fun transforms the SILVA-based Operational Taxonomic Units (OTUs) into a taxonomic profile of KEGG organisms that is normalized by the 16S rRNA copy number (obtained from the NCBI genome annotations) [23]. The result is a table containing relative KEGG Ortholog (KO) abundance levels.

2.3 Phylogenetic transformation of micro biota data for random forest

Most studies of microbiomes analyze the relative abundance of bacterial taxa to make measurements comparable across samples [24]. However, the relative nature of microbial abundance data in microbiota studies can lead to spurious statistical analyses. To avoid spurious statistical analyses because of the relative nature of microbial abundance data in microbiota studies, we used the Phylogenetic Isometric Log-Ratio (PhILR) transformation [19]. The main idea behind the PhILR transformation is to consider the bacterial phylogenetic tree as a natural and informative sequential binary partition to construct an isometric log-ratio that converts compositional data into a real Euclidean space. This phylogenetically driven isometric log-ratio transformation can help to capture the hierarchical pattern of a microbial community structure.

2.4 Pre-screening algorithm for random forest coupled with forward variable selection

We used the Boruta algorithm [15] as a relevant embedded feature selection algorithm that uses the RF classifier to detect all strongly and weakly relevant OTUs (or phylogenetic internal nodes, or functional profiles) to reduce the considerable data dimensionality. This improved the classification accuracies and significantly decreased the time computation. The main idea of this algorithm was to duplicate each OTU, thus creating “shadow” OTUs by randomly permuting the observations of duplicated OTUs at the first step. Then, the importance of all the OTUs is computed (calculated as Z-scores) and the maximum Z-score among the shadow bands is identified when the RF classifier is run. The number of times that the importance of an OTU is higher than the maximum Z-score among the shadow OTUs is counted. An OTU is deemed “important” when the frequency is significantly higher than the expected value, otherwise the OTU is deemed “unimportant” and removed. The global framework of our new approach is shown in (Figure S12).

2.5 Two types of empirical datasets

To examine the performance of our RF-FVS approach, we analyzed two types of empirical datasets from large-scale case control studies as follows.

2.5.1. 16S rRNA gene amplicon dataset

We collected a published case-control 16S rRNA gene amplicon sequencing gut microbiome dataset that included disease meta data and sequencing data with 3347 OTUs for Clostridioides difficile infection (CDI) [17] from 338 individuals; 89 with CDI (cases), 89 with diarrhea who tested negative for CDI (diarrheal controls), and 155 non-diarrheal controls.

2.5.2. Shotgun metagenomics dataset

We collected a published fecal shotgun metagenomic dataset [18] that included 290 samples from tumor-free controls and 285 samples from individuals with CRC. Unlike the 16S rRNA gene amplicon data, shotgun metagenomics provide insight into both microbial community structure and the functions encoded by genomes of the microbiota. For example, the thousands of new protein families that represent novel functions specific to given environments [25], and taxa and the precise dysfunctions of microbial metabolism in gastrointestinal microbiome associated healthy and diseased humans [26] are identified by the full analysis of metagenomic databases.

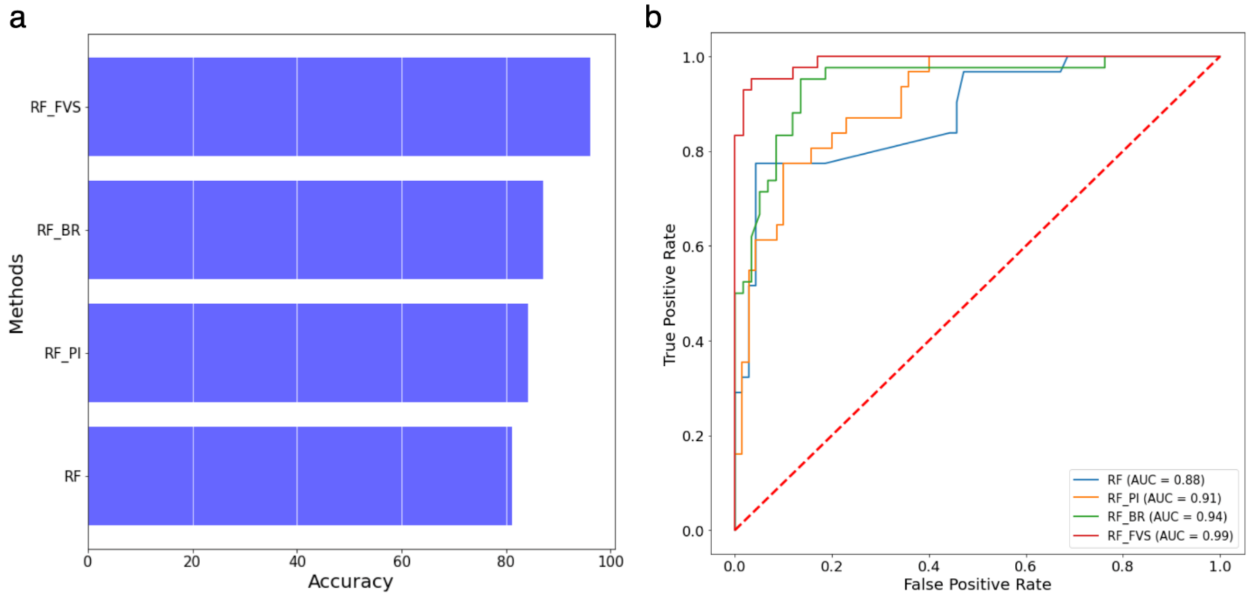

Figure 2: Performance of the random forest classifier for 16S rRNA gene amplicon data from the Clostridioides difficile infection (CDI) dataset. **(a)** Accuracy of random forest classifier with different variable selection methods for the 16S rRNA gene amplicon dataset. **(b)** ROC and AUC of random forest classifier for the 16S rRNA gene amplicon dataset; RF: Random forest algorithm, RF-PI: Random forest algorithm with permutation importance algorithm, RF-BR: Random forest algorithm with Boruta algorithm, RF-FVS: Random forest algorithm with forward variable selection algorithm.

2.6 Software implementation

The RF-FVS framework includes two core modules. First of all, the random forest classifier, permutation importance algorithm, Boruta algorithm and forward variable selection algorithm are implemented by Python programming language in order to optimize the parallel computations that could reduce significantly the computation time. A Python library Scikit-learn is used to implement core computational techniques for the random forest classifier [27]. Secondly, the phylogenetic analysis and visualizations are implemented by using some R packages. For example, the phylogenetic isometric log ratio transformation is implemented in the package philr [19], functional microbiome analysis is implemented in package Tax4Fun2 [21] and phylogenetic visualization is implemented in package ape [28]. Besides, if the phylogenetic tree is not available, the standard phylogenetic analysis methods will be introduced for users such as MEGA [29], RAxML [30] and IQ-TREE [31] in order to build the phylogenetic tree.

3 Results

3.1 Improved accuracy for relative OTU abundance data from the 16S rRNA gene amplicon dataset

We compared the predictive power of our RF-FVS approach to classify three groups (CDI case, diarrheal control, non-diarrheal control) between the micro biome data and the clinical data. The clinical data included age, sex, ethnicity, antibiotic use, antacid use, a vegetarian diet, surgery within the past 6 months, a history of CDI, residence with another person who had CDI, and residence with another person who works in health care. The clinical data were 51% accurate (AUC = 0.7), whereas the microbiome data were 81% accurate (AUC = 0.95) (Figure S1). In particular, the accuracy was high for the microbiome data of the non-diarrheal control group. Moreover, when treating the composition of the microbiota data by Phylogenetic Isometric Log-Ratio (PhILR) transformation, we found that the RF classifier achieved high accuracy (85%; AUC = 0.95). However, it was still difficult to distinguish CDI cases and diarrheal controls in both the microbiome and phylogenetic transformation data. Overall, the accuracy of diagnosis was significantly improved with the RF-FVS approach for the CDI and diarrheal control groups for the microbiome data compared with the accuracy with the RF-PI and RF-BR algorithms (Figure 2). By focusing on 119 species, the accuracy of the RF classifier increased to 96% (AUC = 0.99) (Figure S1). CDI Cases were identified with an accuracy of 94% (AUC = 0.94) and diarrheal controls were identified with an accuracy of 93% (AUC = 0.97). For the phylogenetic transformation data, the accuracy was 95% (AUC = 0.97) when the FVS approach detected the 36 phylogenetic internal nodes.

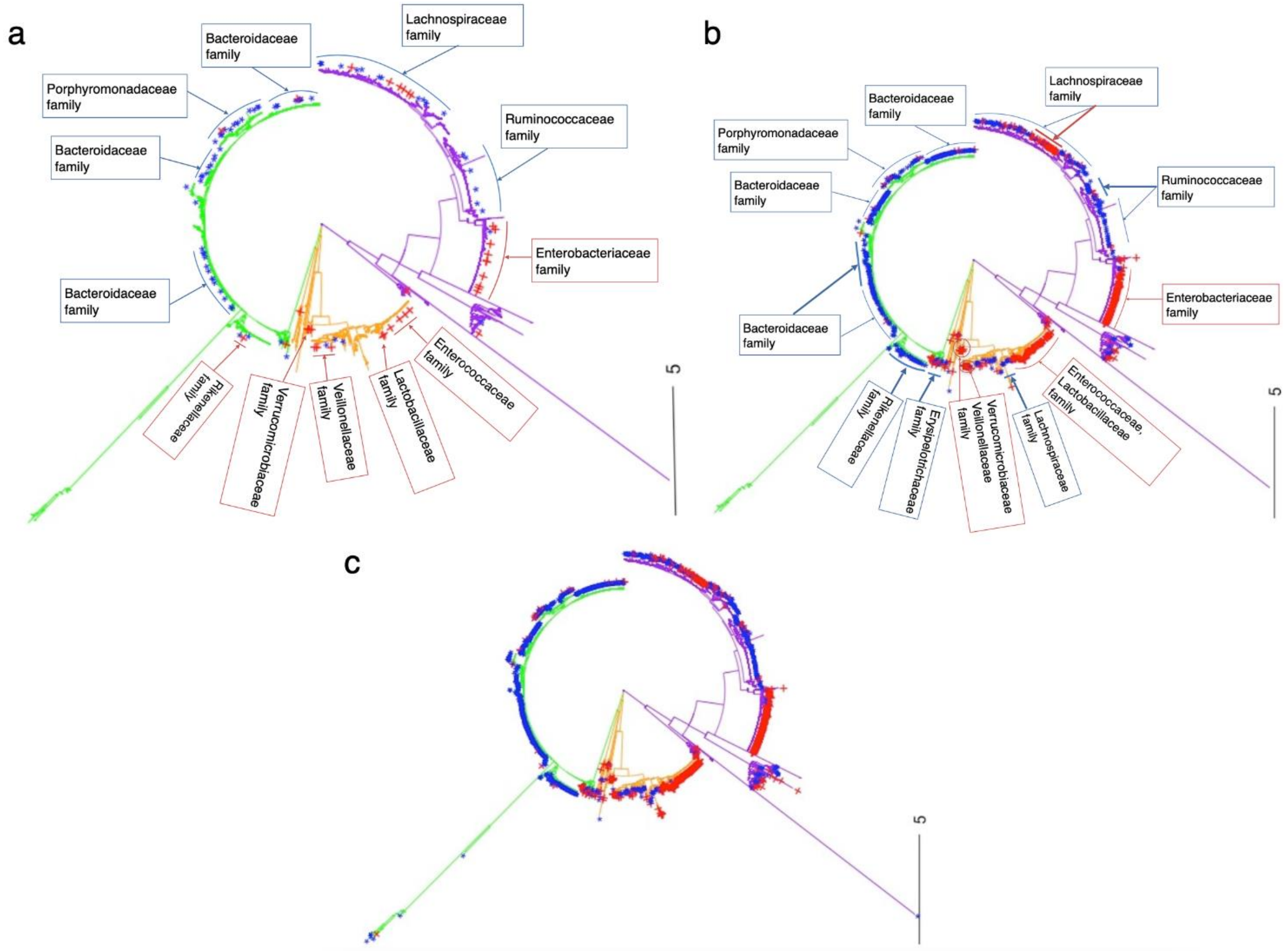

Figure 3: Phylogenetic tree of the 16S rRNA microbiota (OTU) data from the CDI dataset.

(a)Positions indicate the 119 important microbial species that remained after forward variable selection. (b)Positions indicate about 1000 important microbial species that remained after applying the Boruta algorithm. (c)Positions indicate about 2000 important microbial species that remained after applying the permutation importance algorithm. Red indicates OTUs associated with CDI cases; blue indicates OTUs associated with the non-diarrheal controls; green indicates group A; orange indicates group B; purple indicates group C.

3.2 Mapping the selected species on the 16S rRNA phylogenetic tree

The forward variable selection approach avoided the sparse problems in selection of microbial species that the Boruta and permutation importance algorithms could not overcome (Figure 3). Therefore, the number of selected species was significantly smaller than the numbers with the other methods, and the performance of the RF algorithm was improved and easier to interpret. The 119 selected OTUs from nine families were clustered in the 16S rRNA tree: Peptostreptococcaceae, Verrucomicrobiaceae, Veillonellaceae, Ruminococcaceae, Lachnospiraceae, Bacteroides, Porphyromonadaceae, Lactobacillaceae, and Enterobacteriaceae (Figure 3a). They formed three main groups of species that were associated with the differentiation of CDI cases from the non-diarrheal controls. The species in group B in Figure 3a (such as denovo 54, 1326, 601, 2346, and 3486) that showed strong positive correlations with CDI cases belonged to a diverse number of families (Figure S2 and Table S1) such as Peptostreptococcaceae [32], Lactobacillaceae (including Lactobacillus genus), Enterococcaceae (including Enterococcus genus) [33, 34], Verrucomicrobiaceae, and Veillonellaceae [35]. For example, Pérez-Cobas et al. [36] reported that the most striking changes in the microbiome of CDI cases occurred in the Lactobacillaceae family, whose frequency increased from <1% at the beginning of antibiotic treatment to 83.3% and 70% on days 35 and 38 of an antibiotic course, then reduced to 15.5% after antibiotic therapy. The species in group B that were selected by the Boruta and permutation importance algorithms, were similar to those selected using the forward variable selection approach (Figure 3b and 3c).

Most of the species in group C (such as denovo 127, 1399, and 788) that showed positive correlations with CDI cases belonged to the Enterobacteriaceae family (Figure 3a and Table S1). Studies [37-39] have shown that relative overgrowth of members of the Enterobacteriaceae family was one of the main causes of significantly disturbed microbiota in CDI. Thus, C. difficile colonization may be facilitated by increased endotoxin production with increased intestinal permeability. However, in group C, the Boruta and permutation importance algorithms selected more species associated with CDI than the forward variable selection approach (Figure 3b and 3c) because these two algorithms used the decrease of Gini impurity after a node split as the main input data for computational processes in order to select the main features. The corresponding species that became the potential candidates of these two algorithms showed large decreases of impurity after certain split. The reduction of impurity of species became very slow and the differences of impurity between species in the same families were insignificant in the high-dimensional sparse microbial data (Figure S3 and Table S1).

Therefore, the abilities of the Boruta and permutation importance algorithms were influenced significantly even if the statistical tests were used. For example, in group C, the FVS algorithm selected a few species in Lachnospiraceae family that had positively associated OTUs (Figure 3a), but the other algorithms kept a large number of the species in this family that showed poor positive correlations with CDI cases (Figure 3b, 3c and Table S2). In groups B and C, a large number of species in the Lachnospiraceae family that were kept by the Boruta algorithm, were associated with non-diarrheal controls (Figure 3b) but showed poor correlations with non-diarrheal controls (Table S2). Besides, a number of species in the Ruminococcaceae family that were selected only by the Boruta and permutation importance algorithms, showed poor positive correlations with non-diarrheal controls (Figure 3b, 3c and Table S3).

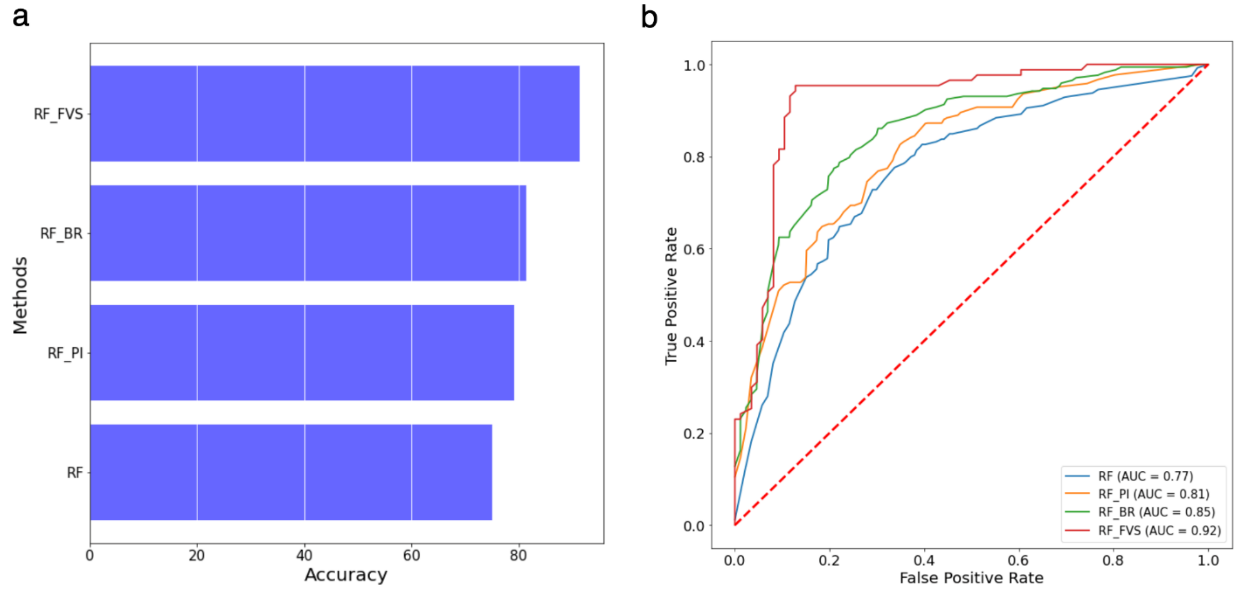

Figure 4: Performance of the random forest classifier for the shotgun metagenomics data of colorectal cancer (CRC) dataset.

(a) Accuracy of random forest classifier with the different variable selection methods for the shotgun metagenomics dataset; **(b)** ROC and AUC of random forest classifier for the shotgun metagenomics dataset. RF: Random forest algorithm, RF-PI: Random forest algorithm with permutation importance algorithm, RF-BR: Random forest algorithm with Boruta algorithm, RF-FVS: Random forest algorithm with forward variable selection algorithm.

Most species in group A (such as denovo 557, 302, 1888, 1987, 1983, and 610), some species in group C (such as denovo 26, 9, 1295, 447, and 347) and one species in group B (denovo 156) that were enriched in the non-diarrheal controls belonged to the Ruminococcaceae, Lachnospiraceae, Bacteroides, and Porphyromonadaceae families (Figure 3a, Figure S2 and Table S1). Short-Chain Fatty Acid (SCFA) production is known to play a principal role in the regulation of intestinal inflammatory processes [40] and intestinal barrier maintenance [41]. CDI was found to cause significant reductions in the Ruminococcaceae and/or Lachnospiraceae families that produce butyrate and SCFA [42, 43]. Moreover, the four species in group A (denovo 2019, 1773, 3205, and 2528) (Figure 3a and Table S1) showed weak associations with CDI case. However, a significantly larger number of species, that had positive correlations with CDI, were selected by the Boruta and permutation importance algorithms (Figure 3b and 3c). For example, in group A, the number species in the Bacteroidaceae family that were kept only by these algorithms, showed insignificant positive correlations with non-diarrheal controls (Table S4). Moreover, the Boruta and permutation importance algorithms selected a huge number of species in the Rikenellaceae and Erysipelotrichaceae families that were ignored by the FVS approach (Figure 3); however, they had very poor correlations with CDI case and non-diarrheal controls (Table S5 and S6). When we applied our RF classifier to the clinical data, we found that antibiotic treatment contributed significantly to an increase in the accuracy of the prediction models [36, 44].

3.3 Improved accuracy for relative OTU abundance data from shotgun metagenomics data

Our RF classifier successfully distinguished the CRC cases and tumor-free controls in both the 16S rRNA gene amplicon data and the shotgun metagenomics data (Figure S4). The average accuracy for the 16S rRNA gene amplicon data, which included 18,448 OTUs, was 68.18% (AUC = 0.73). For the shotgun metagenomics data, the accuracies of the RF classifier for CRC cases and tumor-free controls were 70% (AUC = 0.86) and 80% (AUC = 0.86) respectively. The forward variable selection significantly improved the performance of the RF algorithm as shown in Figure 4. Specifically, the forward variable selection detected 75 microbial species (out of 849 species) that were differentially abundant in the CRC microbiome, which increased the accuracy of the RF classifier to 88% (AUC = 0.92) for the CRC cases and to 95% (AUC = 0.92) for the tumor-free controls (Figure S4). These findings are consistent with previous reports of significant enrichment of novel species in the fecal microbiomes of patients with CRC. For example, we detected the three most important species, Parvimonas micra, Flavonifractor plautii, and Gemella morbillorum, that helped to improve the accuracy of our RF classifier (Figure S5). Gupta et al. [45] found that Flavonifractor plautii was associated significantly and enriched in CRC samples of Indian patients. Flavonifractor plautii was linked with the degradation of beneficial anticarcinogenic flavonoids, and this role was strongly correlated with enzymes and modules involved in flavonoid degradation in CRC samples of the Indian patents. In our study, Flavonifractor plautii was significantly associated with CRC samples of cohorts from France, Germany, China, United States, and Austria (Figure S6).

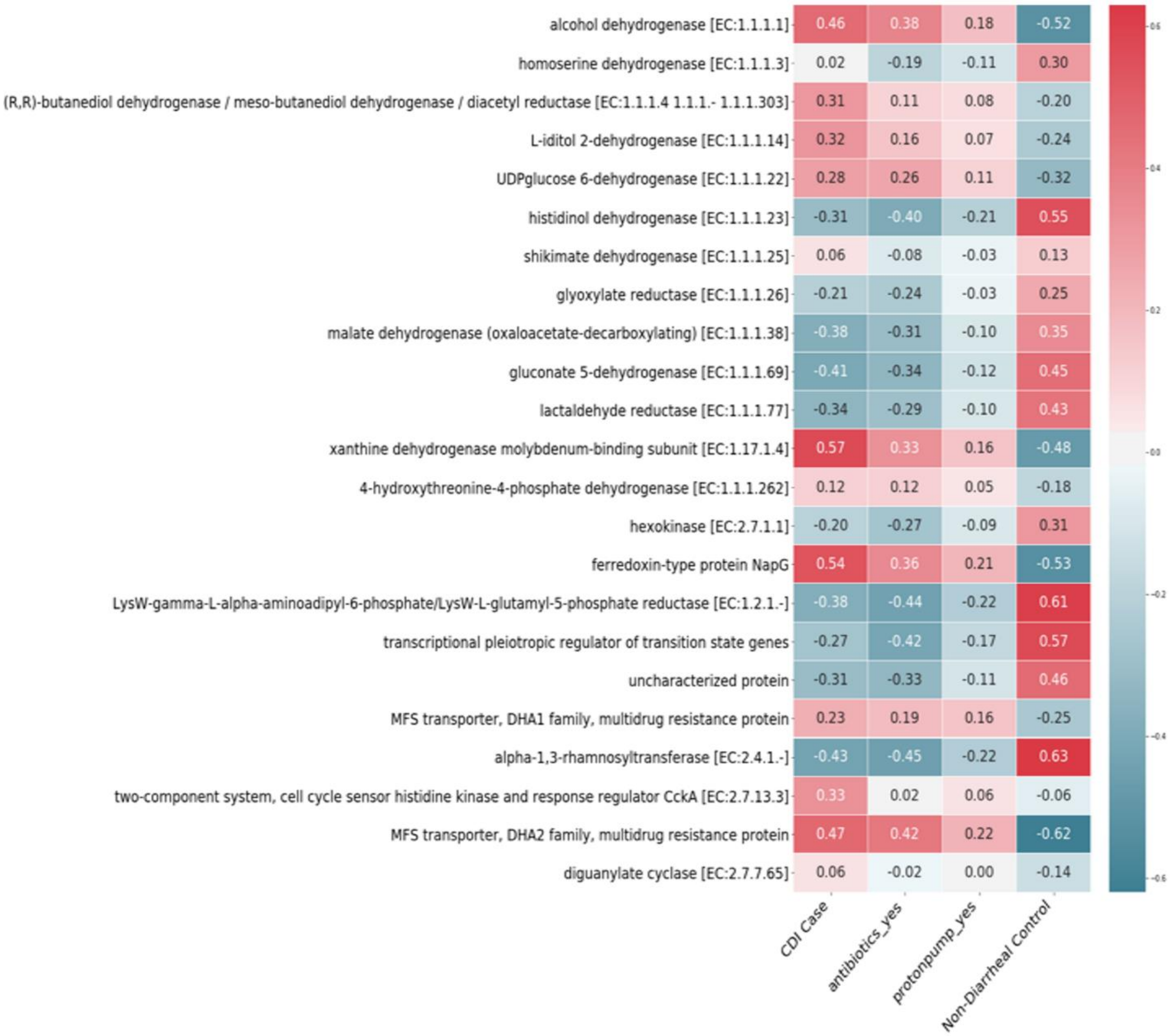

Figure 5: Correlations among the 23 main functional profiles with CDI cases, antibiotics, proton pump inhibitors, and non-diarrheal controls.

3.4 Influential bacterial functions predicted from the 16S rRNA microbiota (OTU) data

The forward variable selection algorithm detected 119 OTUs out of 3347 OTUs. To predict functional profiles from these OTU candidates, we examined 5818 of the 21,620 functional profiles of the KEGG organisms in the Tax4Fun framework. The RF method gave an average accuracy of 81% (AUC = 0.93) for the 5818 predicted functional profiles. Specifically, the accuracies for the CDI cases, diarrheal controls, and non-diarrheal controls were 68%, 79%, and 93% (AUCs = 0.88, 0.92, and 0.97) respectively. To reduce the computational burden, we used the prescreen algorithm at the first step, which reduced the predicted functional profiles from 5818 to 2534. Then, we applied the RF-FVS algorithm, which identified 23 functional profiles (out of 2534) that were different for the CDI cases compared with the controls, which significantly increased the average accuracy to 90% (AUC = 0.95) (Figure S7). Specifically, the accuracies for CDI cases, diarrheal controls, and non-diarrheal controls were 77%, 93%, and 98% (AUCs = 0.90, 0.95, and 0.97) respectively. Some of the 23 most significant functional profiles were strongly associated with CDI cases but a larger number of them were associated with healthy gut (non-diarrheal controls) (Figure 5). Our results confirmed that bacteria support human health with functions such as histidinol dehydrogenase, gluconate 5-dehydrogenase, lactaldehyde reductase, and alpha-1,3-rhamnosyltransferase [46-48]. However, these functions were absent in the microbiomes of patients with CDI, mainly because they were treated with antibiotics and proton pump inhibitors, which killed these bacteria (Figure 5). Therefore, C. difficile, which is resistant to these treatments, became dominant and increased the risk of CDI.

3.5 Influential bacterial functions predicted using the shotgun metagenomics data

We used the evolutionary genealogy of genes from the Non-supervised Orthologous Groups (eggNOG) orthologous gene family abundances and KEGG module abundance profiles to detect functional profiles in the shotgun metagenomics data for CRC. Because the numbers of functions in the KEGG and eggNOG databases were very large (7955 and 31,185 respectively), we used the prescreen algorithm before applying forward variable selection. Our RF-FVS algorithm identified 29 out of 7955 functions in the KEGG database that significantly improved the performance of the RF classifier. Specifically, the accuracy increased from 60% (AUC = 0.79) to 84% (AUC = 0.87) for CRC cases and from 80% (AUC = 0.79) to 97% (AUC = 0.87) for tumor-free controls (Figure S8). A number of functions such as HOMODA hydrolase (K10623), carbamoyl-phosphate synthase 1 (K01948) had strong positive correlations with CRC cases (Figure S9). The contributions of some of these functions to the stage progress of CRC have been reported. For example, carbamoyl-phosphate synthase 1, a metabolic enzyme that utilizes ammonia to produce carbamoyl phosphate, is encoded by one of four novel driver genes that were identified as hubs for stage-III progression of colorectal cancer [49]. Our RF-FVS algorithm also identified 53 out of 31,185 functions in the eggNOG database that significantly improved the performance of the RF classifier. Specifically, the accuracy increased from 62% (AUC = 0.78) to 86% (AUC = 0.90) for CRC cases and from 80% (AUC = 0.78) to 97% (AUC = 0.90) for tumor-free controls (Figure S10). Although a large number of functions were significantly positively correlated with CRC cases, such as ENOG410Y6BY, ENOG410XYS8, ENOG411EMB, and ENOG410ZGTS (Figure S11), experimental information about their functions is lacking. These genes are likely to be good candidates for further studies.

Data type	**Analysis framework**			**Accuracy**	**Time computation**
	Boruta alogirithm	**Random forest**	**Forward algorithm**		
OTUs data	✖	✔	✔	96.04%	1 week
	✔	✔	✔	99.01%	1 day
PhILR transformation	✖	✔	✔	95.05%	26 hours
	✔	✔	✔	95.05%	8.75 hours
Functional profile data	✖	✔	✔	90.10%	4 days
	✔	✔	✔	90.10%	13 hours

Table 1: Random forest classifier with forward variable selection (RF-FVS) with and without the prescreen algorithm for the CDI dataset. ✔, the prescreen algorithm was used; ✖, the prescreen algorithm was not used. All algorithms were run in a parallel environment. The properties of the parallel version were evaluated on a high-performance computer (Intel® Xeon® Gold 6230 Processor 2.10 GHz × 2, 40 cores, 2 threads per core, 93.1-Gb RAM) under Ubuntu 20.04.1 LTS.

3.6 Forward variable selection and the prescreening algorithm reduced the CPU time

Although the RF classifier achieved high accuracy in analyzing the microbiome data, the high dimensionality of the data meant the computational burden was high. The RF classifier took about one week to identify 119 species out of 3347 species in the 16S rRNA gene amplicon data for CDI. By using Boruta algorithm to identify a small number of informative variables, the 3347 species were reduced to 1008 species and the RF-FVS algorithm detected 96 species with an increased accuracy of 99% (AUC = 0.99) (Table 1). The computation time also was reduced from one week to one day. For the functional profile predictions, the FS-FVS algorithm took about 4 days to identify 65 out of 5818 functional profiles. The Boruta algorithm reduced the total number of functional profiles from 5818 to 2534 and the RF-FVS then needed only 13 hours to detect 23 out of the 2534 functional profiles. The accuracy of the RF classifier increased to 90.1%.

4. Discussion

A number of publicly available databases contain information about microbial species and their functions that are associated with disease or health. The large number of microbial species and functional profiles in these databases significantly negatively influence the power of machine learning classifiers, including the RF algorithm. Further, the computational cost of running these algorithms to detect a few informative features in the high-dimensional space of microbiome data is still very high. In this study, we developed a novel procedure that significantly enhances the RF classifier and substantially improves its performance in terms of high-speed computation and high accuracy. We tested its performance using two microbiome datasets that contained a large number of species and functional signatures (>30,000 variables) but a very small proportion of significant variables. Our RF-FVS approach was useful in several respects. Firstly, the RF-FVS algorithm identified the core set of microbial species and their functional profiles, which considerably increased the predictive accuracy of the RF classifier. The highest increase in the predictive accuracy was to 99% for the CDI cases classification. Moreover, because 16S rRNA sequencing data do not directly provide insights into the functional capabilities of the microbiome community, we integrated the Tax4Fun tool into our pipeline to predict the functional profiles of microbial communities based on the 16S rRNA datasets. Therefore, our RF-FVS approach could detect the minimal-size core set and optimal predictive subset of functional profiles of the selected microbial species and the linkages among them, which will provide insights into the ecological functioning of habitats. Some unknown species and genes within the dedicated taxa and functional profiles were detected in the group of selected features that made meaningful contributions to the predictive performance of the RF classifier. These species and genes are likely to be good candidates for future experimental studies.

Secondly, there are a number of standard approaches that can use 16S rRNA data to infer the metabolic potential of the corresponding microbial species. For example, if the database is annotated by the Greengenes database, PICRUSt can achieve good estimations for the functional potential of microbial communities [50]. Tax4Fun is a good option for data annotated by the SILVA database. In this study, the published databases that we used to checked the performance of the RF-FVS approach adopted the Ribosomal Database Project (RDP) approach to classify the 16S rRNA gene sequences taxonomically [51]. Tax4Fun achieved significantly better quality of predicted functional profiles than PICRUSt; therefore, Tax4Fun was integrated as the main option in our pipeline. In the future, we plan to check the performance of the RF-FVS approach for other databases that provide taxonomy annotations, such as Greengenes, LTP, RDP, and SILVA [52]. Moreover, to overcome the computational burden of high-dimensional data that limits the implementation of existing machine learning approaches, we developed a parallel computational strategy algorithm for handling large-scale problems in the forward variable selection algorithm. This parallel strategy helps to equally divide the computational burden of search processes among processors. Thus, the forward variable selection computation process is completely optimized and parallelized based on data partitioning. In microbiome datasets of tens of thousands of species and their functional profiles, selecting only a few hundred of the most significant samples can be a major problem. Our current strategy is focused on parallelly searching for the features of interest. In the future, the numbers of samples and features in microbiome datasets are likely to explode at a rapid pace. We anticipate that hybrid-partitioning strategies that partition the data both horizontally (over samples) and vertically (over features) will become essential to speed up the computational processes [53, 54].

Acknowledgment

We thank Margaret Biswas, PhD, from Edanz Group (https://en-author-services.edanz.com/ac) for editing a draft of this manuscript.

Competing interests

The authors have declared that no competing interests exist

Author Contributions

Conceptualization: Tung Dang, Hirohisa Kishino

Formal analysis: Tung Dang

Funding acquisition: Hirohisa Kishino

Investigation: Tung Dang, Hirohisa Kishino

Methodology: Tung Dang, Hirohisa Kishino

Software: Tung Dang

Supervision: Hirohisa Kishino

Visualization: Tung Dang, Hirohisa Kishino

Writing–original draft: Tung Dang

Writing–review & editing: Tung Dang, Hirohisa Kishino

Funding

This study was supported by a Grant-in-Aid for Scientific Research (B) 19H04070 from the Japan Society for the Promotion of Science.

References

1. Wu GD, Chen J, Hoffmann C, et al. Linking long-term dietary patterns with gut microbial enterotypes. Science 334 (2011): 105-108.
2. Qin J, Li Y, Cai Z, et al. A metagenome-wide association study of gut microbiota in type 2 diabetes. Nature 490 (2012): 55-60.
3. Arumugam M, Raes J, Pelletier E, et al. Enterotypes of the human gut microbiome. Nature 473 (2011): 174-180.
4. Turnbaugh PJ, Ley RE, Hamady M, et al. The human microbiome project. Nature 449 (2007): 804-810.
5. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal statistical society 57 (1995): 289-300.
6. Hothorn T, Hornik K, Van De Wiel MA, et al. A lego system for conditional inference. The American Statistician 60 (2006): 257-263.
7. Knight R, Vrbanac A, Taylor BC, et al. Best practices for analysing microbiomes. Nature Reviews Microbiology 16 (2018): 410-422.
8. Zhou YH, Gallins P. A review and tutorial of machine learning methods for microbiome host trait prediction. Frontiers in Genetics10 (2019): 579.
9. Breiman L. Random forests. Machine learning 45 (2001): 5-32.
10. Ho TK. The random subspace method for constructing decision forests. IEEE transactions on pattern analysis and machine intelligence 20 (1998): 832-844.
11. Dietterich TG. An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. Machine learning 40 (2000): 139-157.
12. Bureau A, Dupuis J, Falls K, et al. Identifying SNPs predictive of phenotype using random forests. Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society 28 (2005): 171-182.
13. Díaz-Uriarte R, De Andres SA. Gene selection and classification of microarray data using random forest. BMC bioinformatics 7 (2006): 3.
14. Knights D, Costello EK, Knight R. Supervised classification of human microbiota. FEMS microbiology reviews 35 (2011): 343-359.
15. Kursa MB, Rudnicki WR. Feature selection with the Boruta package. J Stat Softw 36 (2010): 1-13.
16. Altmann A, Toloşi L, Sander O, et al. Permutation importance: a corrected feature importance measure. Bioinformatics 26 (2010): 1340-1347.
17. Alyxandria MS, Mary AMR, Cathrin R, et al. Microbiome data distinguish patients with Clostridium difficile infection and non-C. Difficile-associated diarrhea from healthy controls. MBio 5 (2014): 1021-1014.
18. Wirbel J, Pyl PT, Kartal E, et al. Meta-analysis of fecal metagenomes reveals global microbial signatures

that are specific for colorectal cancer. Nature medicine 4 (2019): 679-689.

19. Silverman JD, Washburne AD, Mukherjee S, et al. A phylogenetic transform enhances analysis of compositional microbiota data. Elife 6 (2017): 21887.

20. Aßhauer KP, Wemheuer B, Daniel R, et al. Tax4Fun: predicting functional profiles from metagenomic 16S rRNA data. Bioinformatics 31 (2015): 2882-2884.

21. Wemheuer F, Taylor JA, Daniel R, et al. Tax4Fun2: prediction of habitat-specific functional profiles and functional redundancy based on 16S rRNA gene sequences. Environmental Microbiome 15 (2020): 1-12.

22. Quast C, Pruesse E, Yilmaz P, et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic acids research 41 (2012): 590-596.

23. Kanehisa M, Goto S, Sato Y, et al. Data, information, knowledge and principle: back to metabolism in KEGG. Nucleic acids research 42 (2014): 199-205.

24. Gloor GB, Macklaim JM, Pawlowsky-Glahn V, et al. Microbiome datasets are compositional: and this is not optional. Frontiers in microbiology 15 (2017): 2224.

25. Godzik A. Metagenomics and the protein universe. Current opinion in structural biology 21 (2011): 398-403.

26. Morgan XC, Tickle TL, Sokol H, et al. Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment. Genome biology 13 (2012): 79.

27. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine learning in Python. The Journal of machine Learning research 12 (2011): 2825-2830.

28. Paradis E, Claude J, Strimmer K. APE: analyses of phylogenetics and evolution inR language. Bioinformatics 20 (2004): 289-290.

29. Tamura K, Stecher G, Kumar S. MEGA11: molecular evolutionary genetics analysis version 11. Molecular Biology and Evolution 38 (2021): 3022-3027.

30. Kozlov AM, Darriba D, Flouri T, et al. RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. Bioinformatics 35 (2019): 4453-4455.

31. Minh BQ, Schmidt HA, Chernomor O, et al. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. Molecular biology and evolution 37 (2020): 1530-1534.

32. Milani C, Ticinesi A, Gerritsen J, et al. Gut microbiota composition and Clostridium difficile infection in hospitalized elderly individuals: a metagenomic study. Scientific reports 6 (2016): 1-12.

33. Ross CL, Spinler JK, Savidge TC. Structural and functional changes within the gut microbiota and susceptibility to Clostridium difficile infection. Anaerobe 41 (2016): 37-43.

34. Ling Z, Liu X, Jia X, et al. Impacts of infection with different toxigenic Clostridium difficile strains on faecal microbiota in children. Scientific reports 4 (2014): 7485.

35. De Wolfe TJ, Eggers S, Barker AK, et al. Oral probiotic combination of Lactobacillus and Bifidobacterium alters the gastrointestinal microbiota during antibiotic treatment for Clostridium difficile infection. PLoS One 13 (2018): 0204253.

36. Pérez-Cobas AE, Artacho A, Ott SJ, et al. Structural and functional changes in the gut microbiota associated to Clostridium difficile infection. Frontiers in microbiology 5 (2014): 335.

37. Tanaka S, Kobayashi T, Songjinda P, et al. Influence of antibiotic exposure in the early postnatal period on the development of intestinal microbiota. FEMS

Immunology & Medical Microbiology 56 (2009): 80-87.

38. Prasad N, Labaze G, Kopacz J, et al. Asymptomatic rectal colonization with carbapenem-resistant Enterobacteriaceae and Clostridium difficile among residents of a long-term care facility in New York City. American journal of infection control 44 (2016): 525-532.

39. Seddon MM, Bookstaver PB, Justo JA, et al. Role of early de-escalation of antimicrobial therapy on risk of Clostridioides difficile infection following Enterobacteriaceae bloodstream infections. Clinical Infectious Diseases 69 (2019): 414-420.

40. Maslowski KM, Vieira AT, Ng A, et al. Regulation of inflammatory responses by gut microbiota and chemoattractant receptor GPR43. Nature 461 (2009): 1282-1286.

41. Koruda MJ, Rolandelli RH, Bliss DZ, et al. Parenteral nutrition supplemented with short-chain fatty acids: effect on the small-bowel mucosa in normal rats. The American journal of clinical nutrition 51 (1990): 685-689.

42. Lawley TD, Clare S, Walker AW, et al. Targeted restoration of the intestinal microbiota with a simple, defined bacteriotherapy resolves relapsing Clostridium difficile disease in mice. PLoS Pathog 8 (2012): 1002995.

43. Antharam VC, Li EC, Ishmael A, et al. Intestinal dysbiosis and depletion of butyrogenic bacteria in Clostridium difficile infection and nosocomial diarrhea. Journal of clinical microbiology 51 (2013): 2884-2892.

44. Theriot CM, Koenigsknecht MJ, Carlson JrPE, et al. Antibiotic-induced shifts in the mouse gut microbiome and metabolome increase susceptibility to Clostridium difficile infection. Nature communications 5 (2015): 3114.

45. Gupta A, Dhakan DB, Maji A, et al. Association of Flavonifractor plautii, a flavonoid-degrading bacterium, with the gut microbiome of colorectal cancer patients in India. mSystems 4 (2019): 438-519.

46. Agyirifo DS, Wamalwa M, Otwe EP, et al. Metagenomics analysis of cocoa bean fermentation microbiome identifying species diversity and putative functional capabilities. Heliyon 5 (2019): 02170.

47. Xie M, Wu J, An F, et al. An integrated metagenomic/metaproteomic investigation of microbiota in dajiang-meju, a traditional fermented soybean product in Northeast China. Food Research International 115 (2019): 414-424.

48. O'Callaghan A, van Sinderen D. Bifidobacteria and their role as members of the human gut microbiota. Frontiers in microbiology 7 (2016): 925.

49. Palaniappan A, Ramar K, Ramalingam S. Computational identification of novel stage-specific biomarkers in colorectal cancer progression. PloS one 11 (2016): 0156665.

50. Langille MG, Zaneveld J, Caporaso JG, et al. Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. Nature biotechnology 31 (2013): 814-821.

51. Wang Q, Garrity GM, Tiedje JM, et al. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Applied and environmental microbiology 73 (2007): 5261-5267.

52. Edgar R. Taxonomy annotation and guide tree errors in 16S rRNA databases. Peer J 6 (2018): 5030.

53. Xing EP, Ho Q, Xie P, et al. Strategies and principles of distributed machine learning on big data. Engineering 2 (2016): 179-195.

54. Lee S, Kim JK, Zheng X, et al. On model parallelization and scheduling strategies for distributed machine learning. In Advances in neural information processing systems 4 (2014): 2834-2842.